WHAT IS CLAIMED:

1. A method of identifying semantic units within a search query comprising:

identifying documents relating to the query by comparing search terms in the query to an index of a corpus;

generating a plurality of substrings from the query;

calculating, for each of the generated substrings, a value that corresponds to a comparison between one or more of the identified documents and the generated substring; and

selecting semantic units from the generated substrings based on the calculated values.

2. The method of claim 1, wherein the identification of the documents further includes:

generating an initial list of relevant documents; and

selecting a predetermined number of most relevant ones of the documents in the initial list as the identified documents.

3. The method of claim 1, wherein the selection of the semantic units further includes:

selecting semantic units from the generated substrings that have calculated values above a predetermined threshold.

4. The method of claim 3, wherein the selection of the semantic units further includes:

discarding the generated substrings that overlap other ones of the generated substrings with higher calculated values.

5. The method of claim 1, wherein the calculated values are weighted based on a ranking defined by relevance of the identified documents, such that substrings that occur in more relevant ones of the identified documents are assigned higher calculated values than substrings that occur is less relevant ones of the documents.

6. A method of locating documents in response to a search query, the method comprising:

receiving the search query from a user;

generating a list of relevant documents based on search terms of the query;

identifying a subset of documents that are most relevant ones of the documents in the list of relevant documents;

generating a plurality of substrings of the query;

calculating, for each of the generated substrings, a value related to one or more documents in the subset of documents that contain the substring;

selecting semantic units from the generated substrings based on the calculated values; and

refining the generated list of relevant documents based on the selected semantic units.

7. The method of claim 6, wherein the identified subset includes a predetermined number of the most relevant ones of the documents in the list of relevant documents.

8. The method of claim 6, wherein the selection of the semantic units further includes:

selecting semantic units from the generated substrings that have calculated values above a predetermined threshold.

9. The method of claim 8, wherein the selection of the semantic units further includes:

discarding the generated substrings that overlap other ones of the generated substrings with higher calculated values.

10. The method of claim 6, wherein the calculated values are weighted based on a ranking defined by relevance of the identified documents, such that substrings that occur in more relevant ones of the documents are assigned higher calculated values than substrings that occur is less relevant ones of the documents.

11. A system comprising:

a server connected to a network, the server receiving search queries from users via the network, the server including:

at least one processor; and

a memory operatively coupled to the processor, the memory storing program instructions that when executed by the processor, cause the processor to: identify a list of documents relating to the search query by matching individual search terms in the query to an index of a corpus; generate a plurality of substrings from the query; calculate, for each of the generated substrings, a value relating to one or more documents of the identified list of documents that contain the generated substring; and select semantic units from the generated substrings based on the calculated values.

12. The system of claim 11, wherein the processor refines the identified list of documents based on the selected semantic units.

13. The system of claim 12, wherein the system transmits the refined list of documents to the user.

14. The system of claim 11, wherein the network is the Internet and the corpus is a collection of web documents.

15. The system of claim 11, wherein the memory includes instructions for causing the processor to:

select semantic units from the generated substrings that have calculated values above a predetermined threshold.

16. The system of claim 15, wherein the memory includes instructions for causing the processor to:

discard substrings that overlap other substrings with a higher calculated value.

17. The system of claim 11, wherein the calculated values are weighted based on a ranking defined by relevance of the identified documents, such that substrings that occur in more relevant documents are assigned higher calculated values than substrings that occur is less relevant documents.

18. A server comprising:

a processor; and

a memory operatively coupled to the processor, the memory including:

a ranking component configured to return a list of documents ordered by relevance in response to a search query; and

a semantic unit locator component configured to locate semantic units in search queries entered by a user based on a predetermined number of most relevant documents in the list of documents returned by the ranking component.

19. The server of claim 18, further including:

a search engine configured to refine the list of documents based on the located semantic units.

20. The server of claim 19, wherein the processor is configured to:

transmit the refined list of documents to a user that provided the query.

21. The server of claim 18, wherein the semantic unit locator is further configured to:

generate a plurality of substrings of the query;

calculate, for each generated substring, a value relating to the portion of the predetermined number of the most relevant documents that contain the substring; and

locate the semantic units from the generated values.

22. The server of claim 21, wherein the semantic unit locator is configured to locate semantic units from the generated substrings that have calculated values above a predetermined threshold.

23. The server of claim 22, wherein the semantic unit locator is configured to discard substrings that overlap other substrings with a higher calculated value.

24. The server of claim 21, wherein the calculated values are weighted based on a ranking defined by relevance of the identified documents, such that substrings that occur in more relevant documents are assigned higher calculated values than substrings that occur is less relevant documents.

25. A computer-readable medium storing instructions for causing at least one processor to perform a method that identifies semantic units within a search query, the method comprising:

identifying documents relating to the query by matching individual search terms in the query to an index of a corpus;

forming a plurality of substrings of the query;

calculating, for each of the substrings, a value relating to the portion of the identified documents that contain the substring; and

selecting semantic units from the generated substrings based on the calculated values.

26. The computer-readable medium of claim 25, wherein the identification of the set of documents further includes:

generating an initial list of relevant documents; and

selecting a predetermined number of the most relevant documents in the initial list to include in the identified documents.

27. The computer-readable medium of claim 25, wherein the selection of the semantic units further includes:

selecting semantic units from the generated substrings that have calculated values above a predetermined threshold.

28. The computer-readable medium of claim 27, wherein the selection of the semantic units further includes:

discarding substrings that overlap other substrings with a higher calculated value.

29. The computer-readable medium of claim 27, wherein the calculated values are weighted based on a ranking defined by relevance of the identified documents, such that substrings that occur in more relevant documents are assigned higher calculated values than substrings that occur is less relevant documents.

30. A computer-readable medium storing instructions for causing a processor to perform a method, the method comprising:

receiving the search query from a user;

generating a list of relevant documents based on individual search terms of the query;

identifying a subset of documents that are the most relevant documents from the list of relevant documents;

forming a plurality of substrings of the query;

calculating, for each of the substrings, a value related to the portion of the subset of documents that contain the substring;

selecting semantic units from the generated substrings based on the calculated values; and

refining the generated list of relevant documents based on the selected semantic units.

31. The computer-readable medium of claim 30, wherein the identified subset includes a predetermined number of the most relevant documents from the list of relevant documents.

32. The computer-readable medium of claim 30, wherein the selection of the semantic units further includes:

selecting semantic units from the generated substrings that have calculated values above a predetermined threshold.

33. The computer-readable medium of claim 32, wherein the selection of the semantic units further includes:

discarding substrings that overlap other substrings with a higher calculated value.

34. The computer-readable medium of claim 30, wherein the calculated values are weighted based on a ranking defined by relevance of the identified documents, such that substrings that occur in more relevant documents are assigned higher calculated values than substrings that occur is less relevant documents.

35. The computer-readable medium of claim 30, wherein the computer-readable medium is a CD-ROM, floppy disk, tape, flash memory, system memory, hard drive, or data signal embodied in a carrier wave.

36. An apparatus for locating documents in response to a search query, comprising:

means for receiving the search query from a user;

means for generating a list of relevant documents based on individual search terms of the query;

means for identifying a subset of documents that are the most relevant documents from the list of relevant documents;

means for forming a plurality of substrings of the query;

means for calculating, for each of the substrings, a value related to the portion of the subset of documents that contain the substring;

means for selecting semantic units from the generated substrings based on the calculated values; and

means for refining the generated list of relevant documents based on the selected semantic units.